Kadyan, Virender; Kathania, Hemant; Govil, Prajjval; Kurimo, Mikko

Synthesis Speech Based Data Augmentation for Low Resource Children ASR

Published in:
Speech and Computer - 23rd International Conference, SPECOM 2021, Proceedings

DOI:
10.1007/978-3-030-87802-3_29

Published: 01/01/2021

Document Version
Peer reviewed version

Please cite the original version:
Kadyan, V., Kathania, H., Govil, P., & Kurimo, M. (2021). Synthesis Speech Based Data Augmentation for Low Resource Children ASR. In A. Karpov, & R. Potapova (Eds.), *Speech and Computer - 23rd International Conference, SPECOM 2021, Proceedings* (pp. 317-326). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 12997 LNAI). https://doi.org/10.1007/978-3-030-87802-3_29

Synthesis speech based data augmentation for low resource children ASR

Virender Kadyan[1,3], Hemant Kathania[2,3], Prajjval Govil[3], and Mikko Kurimo[3]

[1] Speech and Language Research Centre, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India
vkadyan@ddn.upes.ac.in
[2] Department of Electronics and Communication Engineering, National Institute of Sikkim, India
hemant.kathania@aalto.fi and hemant.ece@nitsikkim.ac.in
[3] Department of Signal Processing and Acoustics, Aalto University, Finland
prajjval.govil@aalto.fi and mikko.kurimo@aalto.fi

Abstract. Successful speech recognition for children requires large training data with sufficient speaker variability. The collection of such a training database of children's voices is challenging and very expensive for zero/low resource language like Punjabi. In this paper, the data scarcity issue of the low resourced language Punjabi is addressed through two levels of augmentation. The original training corpus is first augmented by modifying the prosody parameters for pitch and speaking rate. Our results show that the augmentation improves the system performance over the baseline system. Then the augmented data combined with original data and used to train the TTS system to generate synthesis data and extended dataset is further used for augmented by generating children's utterances using text-to-speech synthesis and sampling the language model with methods that increase the acoustic and lexical diversity. The final speech recognition performance indicates a relative improvement of 50.10% with acoustic and 57.40% with language diversity based augmentation in comparison to that of the baseline system respectively.

Index Terms: Low resource, children speech recognition, prosody modification, speech synthesis, tacotron

1 Introduction

In recent years, Automatic speech recognition (ASR) system for children speech is an active area of research. It also showed a significant progress on inclusiove ASR technologies. At the same time, spoken interfaces are utilized increasingly in smart devices through applications like Amazon Alexa, Google Home and Apple Siri [1]. The wide and successful real world use of such a general purpose spoken interface requires a vast amount of training data for the ASR system [2]. Still, only a few languages like English, Chinese and Arabic have sufficient speech and language data resources and, in fact, most other languages spoken

in the world have low or even zero resources. Research programs like IARPA Babel have facilitated ASR development in many low resourced languages [3], but there are still languages like Punjabi spoken in Indian part of Punjab where speech resources for children are zero and for adults very small [1] even though it is spoken by as many as 105 million native speakers.

For building large vocabulary ASR system large amount of training data is required. This challenge occurs because of the high necessity of acoustic, speaker and linguistic variability and the cost of the large-scale human transcription work. In the past, efforts have been made to explore training data of other languages to be employed via multilingual deep neural networks. Additional corpora have been generated through in or out domain strategies using transfer learning [4], model-agnostic meta learning algorithm [5], and triangular architecture [6] etc. They contributed towards higher recognition rate in various other languages [7]. A few studies on low resource languages have investigated whether limited training speakers could be extended using single or multiple data augmentation [3, 8, 9]. Recently various data augmentation approaches have been explored using language adaptive DNN [10], mapping of acoustic feature through conversion of adult voice into child voice, processing of corrupted clean data [11, 12], and speech dereverberation [13] are simulated using Generative adversarial network (GAN) [1], speed, volume and spectogram perturbation. These approaches works through masking of time or frequency parameters [14].

So far, few general purpose ASR has focused upon data augmentation using the training speakers only. This can be performed, for example, by prosody modification of the training data [15, 16]. This augmentation approach is not much beneficial for ASR system. It occurred due to lack of robustness in parameter estimation. Consequently, prosody transfer based ASR-TTS methods are also found to be effective in generation of audios through voice conversion [17].

Till date very limited work on text-to-speech synthesis (TTS) have been employed for low resource languages [18]. A few other voice mapping techniques between train or test utterances have been performed by Tacotron 2[19], Transformer TTS [20] and Fast Speech [21]. Unfortunately, on small training sets this generates unnatural voices which are not beneficial for training the ASR system. A range of approaches have been examined to overcome such issues. These involve either tuning the original voice or generation of utterances from the text data. This has recently created a lot of attention by cascading ASR-TTS for generation of synthetic copies. This behaves like a recognition-conversion system which carries language specific information [22, 23]. The task of generation of audios from text is leveraged using TTS synthesis [24]. For most languages it is easier to find or prepare text than speech corpora. A TTS helped in replication of its training data. It contain the characteristics which is similar to the training of an ASR model. However, this approach is well suited for low resource languages where there is no training data. Previous studies on data augmentation mostly focused on individual approaches only. Not much work has been reported

[1] LDC-IL, Punjabi Raw Speech Corpus, https://data.ldcil.org/punjabi-raw-speech-corpus last accessed 2021/05/10

on examining the effect of TTS under low resource circumstances. Though this study includes cascading of TTS-ASR for generation of synthetic data not only on original speech but also indulge the characteristics of prosody data which provide speech which is equivalent to that of human original audios quality.

In this paper, we first collected new speech data for Punjabi children, and created a baseline ASR system using TDNN acoustic model in Kaldi. However, the baseline result was poor because of the limited amount of training data. To capture more acoustic and speaker variability to alleviate the data scarcity we tried two types of data augmentation methods. Initial augmentation was performed by modifying prosody parameters (pitch and speaking rate) to introduce more acoustic variability. Further we used the prosody based augmented data to generate synthetic speech data using a Tacotron-based TTS system. In the TTS system we investigated acoustic and lexical diversity methods in data generation. We also explored sampling new sentences from our language model in addition to both the diversity models.

2 Data

In this study, a Punjabi language children speech corpus has been collected from native speakers of India part of Punjab. Despite being spoken or written by large number of speakers it is still declared as under-resource language. Our self created corpus was built through read speech. The Punjab School Education Board books were read aloud by pupils from 7 to 14 years. A few pupils also contributed spontaneous speech by explaining on a certain topic like about themselves, their community etc. The corpus was collected from 79 speakers (35 male and 44 female) which were further divided into train and test speakers in 80:20 ratio. No speakers nor sentences occur both in the train and test data. Corpus statistics like age group, duration etc are presented in Table 1. The collected speech were varying with different prosody parameters like speaking rate (SR), and pitch. It resulted into huge variations among speech which were collected from different speakers. Pupils up to 10 years old spoke long sentences by concatenating or by taking a short pause between two continuous sentences but above 10 years old pupils spoke sentences fluently. The corpus was recorded through mobile devices in a controlled noise-free environment. The speech data was sampled at 16 kHz. Each speaker spoke long sentences of up to 8-9 words.

3 Baseline system

We used a Kaldi toolkit-based recipe to train the baseline ASR system [25]. It utilizes conventional MFCC features using 40 channel Mel-filterbank with a frame size of 25ms and frame shift of 10ms to train DNN and TDNN-based acoustic models [26, 27]. For normalization, cepstral feature-space maximum likelihood linear regression (fMLLR) was used. The fMLLR transformations for the training and test data were generated using the speaker adaptive training [28]. LDA-MLLT+SAT based GMM alignment labels were used to train the DNN and

Table 1. Indian Punjabi speech corpus details

Purpose	Training	Testing
No. of speakers	63 (28 m and 35 f)	16 (7 m and 9 f)
Speaker age	7-14 years	7-14 years
No. of sentences	8420	2250
Duration (hrs.)	12.20	2.50
Vocabulary Size	8587	2588

TDNN acoustic models. The initial as well as final learning rates were kept as 0.005 and 0.0005, respectively. Further Tanh nonlinearity was used in the DNN architecture. The number of hidden nodes was kept at 512 whereas a total of 3 hidden layers are employed in it. On the other hand, i-vector [29] based speaker adaptation was used for the TDNN based acoustic model. The decoding is then performed using a bigram language model for the DNN system and a 4-gram maximum entropy language model built using SRILM toolkit [30] for the TDNN system. Baseline WER's for the DNN and TDNN -based systems are given in Table 2. From Table 2 it can be noted that the TDNN system outperforms the DNN so for further study we used the TDNN system.

Table 2. WER obtained on the baseline DNN and TDNN acoustic models

Acoustic model	WER (%)
DNN	12.73
TDNN	9.18

4 Data Augmentation

In this paper, we tried to implement three type of augmentation approaches like prosody based data augmentation (Section 4.1), TTS based data augmentation (Section 4.2) and language model sampling based data augmentation (Section 4.3).

4.1 Prosody based data augmentation

We change two types of prosody parameters, the pitch scale (PS) and the speaking rate (SR), systematically to add more prosody variation in the children's

speech (Section 2). We then augment the modified data to the original corpora for further system development. To modify the pitch and speaking rate, we have explored Time Scale Modification (TSM) based on Real-Time Iterative Spectrogram Inversion with Look-Ahead (RTISI-LA) algorithm [16, 31]. Both these prosody parameters are tunable and we varied the pitch modification factorr s from 0.65 to 1.45 to modify pitch and the speaking rate modification factor α from 0.65 to 1.85 with a step size of 0.10. The best values of pitch 0.85 and speaking rate 1.15 were selected on the lower WER of test set and later these values are utilized for further experimentation. RTISI-LA algorithm constructs a high-quality time-domain signal from its short-time magnitude spectrum. The effect of the prosody based data augmentation made to the original data is reported in Table 3. We found that combining the pitch scale and speaking rate based data augmentation improves most over the baseline.

Table 3. WER obtained on prosody modification based data augmented system.

System Type	WER (%)
Original (O)	9.18
O+SR	9.06
O+PS	8.98
O+PS+SR	7.84

4.2 Speech synthesis based data augmentation

In the previous section, we proposed prosody based data augmentation to capture more acoustic variability. The prosody modified data was then combined with the original data to train a TTS model based on Tacotron2[19]. This not only provides the effect of synthesized copies of the original audios, but it also induces characteristics of the prosody modified audios. To produce diverse speaker representation on train utterances we employed an acoustic diversity method which improved the ASR performance over the baseline ASR. Consequently, the impact of lexical diversity for TTS utterances was evaluated using Tacotron 2. Only few synthetic utterances were not found to be useful and others were resemble to natural speakers' audios. Finally we can artificially increase the acoustic and lexical diversity of the training data.

Acoustic Diversity In this section, a Tacotron2 based text-to-speech synthesis model architecture was used to process an input text sequence. It is a combination of encoder-decoder network with an attention mechanism, and a Wavenet based Vocoder. It takes input as a sequence of text in Punjabi language, which is encoded by encoder. In the first part of the encoder, the character sequence

is converted into a word embedding vector. The input text sequence embedding is encoded by 3 convolution layers each containing 512 filters of shape 5 x 1, followed by a bidirectional LSTM layer of 250 units for each direction. Tacotron 2 also uses 'Local sensitive attention' which takes the encoder output as input and tries to summarize the full encoded sequence as a fixed length context vector for each decoder output step. Later, a decoder is employed which is an autoregressive recurrent neural network. It predicts a mel spectrogram from the encoded input sequence one frame at a time. The output of the attention layer is passed through a small pre-net containing 256 hidden ReLU (Rectified Linear Unit) units. The pre-net output and attention context vector are concatenated and passed through a stack of LSTM layers. The output of the LSTM layer is projected through a linear transform to predict the target spectrogram frame. This helps in prediction of a normalized weight vector which can be further employed for aggregation on the basis of attention history. The predicted mel spectrogram is passed through convolution postnet layers. The postnet layer predicts a residual to add to the prediction to improve the overall reconstruction. Finally, the mel spectrogram is transformed into time domain waveforms by modified Wavenet vocoder. The mel spectrograms are mapped to a fixed dimensional embedding vector, known as deep speaker vectors (d-vectors). These d-vectors are frame-level speaker discriminative features that represent the speaker characteristics. The proposed system uses three different approaches to generate d-vectors for inference to handle speaker diversity in the synthesized data. Training of the TTS is done with a GMM based attention model followed by LSTM layers. All the local encoder input is processed and fed to its decoder. This provides data duplicates by synthesized data which was later used for training utterances augmentation. In these, training data was extended by combining one more type of augmented data i.e. synthesized copies which were generated with the help of training utterances. It has been performed by using three different types of speaker conditioning information [32]:

- Original (O'): A set of d-vector is employed as speaker condition information on train data. In this, the output is synthetic copies that resemble to the source data.
- Sampled (S): In this case d-vectors were obtained on random selection of speakers which were different in characteristics to that of source utterances and synthesized utterances. Apart, these speaker information were seen earlier by the synthesizer.
- Random (R): A random set of 256 dimensional vector information were generated after projected them on unit hypersphere through L2 normalization. It was employed further for d-vector and employed in synthesized utterances.

In these types of synthesized dataset, the acoustic diversity (AD) helped in controlling of speaker diversity among the training dataset. It was performed by extracting speaker information as d-vectors. On the basis of O', S, and R synthetic speech, different augmented data has been generated which were further added to original data to train an ASR system. It was generated on basis of

data in train, train+test and train+new speaker based TTS model. The resulting augmented data introduced more speaker variability. As shown in Table 4 the Sampled (S) type gives a lower WER than others and the best performance was achieved by further augmentation of baseline + tacotron (O'+S+R) data.

Table 4. WER obtained on acoustic diversity based augmented ASR system.

AD	WER (%)
None	-
Original (O')	6.69
Random (R)	7.14
Sampled (S)	6.27
O+O'+R+S	5.62

Table 5. WER obtained on lexical diversity based augmented ASR system.

LD	WER (%)
Sampled (S')	4.84
Original (O")	4.79
S'+O"	4.18

Lexical Diversity The purpose of using Lexical Diversity (LD) based TTS synthesis was to generate audio from text transcripts while using the model speaker's characteristics. It helped in production of speech which resulted into addition of lexical as well as acoustic variations in our train data. One approach to add lexical diversity is to use the transcripts in the test data by ignoring the audios corresponding to them. Two types of augmented data were generated as Original (O") and Sampled (S'). In case of S' there was no involvement of test audio whereas in O" the speakers were synthesized from the test speakers' audios. In such case it significantly enhanced the performance of ASR system which would not be generally true in real world situations. However, such LD approach could be beneficial in speaker and domain specific ASR applications. However, the LD results presented in Table 5 show that O" where augmenting was based on knowledge of the test audios is only slightly better than S' which was augmented

with only test transcripts and copies of train utterances. To obtain more gain on ASR performance, the output audios generated from combination of O" + S" systems were further pooled with the baseline.

4.3 Language Model Sampling

To further boost the performance of AD and LD based augmented systems, Language model sampling has been performed with most common lexicon based utterances of Punjabi language that had a maximum length of 15 words. The collected utterances were synthesized and pooled with training set audios to lower the WER of the AD and LD systems. The utterance augmentation was found to be beneficial to some extent as depicted in Tables 6. These systems achieved the lowest WER of 4.58% and 3.91% on AD or LD augmented system. However, the system performance degraded when too much variation of training utterances was augmented as synthetic utterances. The best results were found at augmenting 10K utterances as presented in Table 6.

Table 6. WER obtained on language model sampling augmentation for the baseline ASR system using acoustic diversity (AD) and lexical diversity (LD)

LMS	WER (%)	
	AD	LD
0	5.62	4.18
5K	4.71	4.02
10K	4.58	3.91
25K	4.97	4.17

5 Conclusion

We have collected new speech data for Punjabi children to build ASR system in this language specifically for children's speech. The baseline system built on this data was not good. So to capture more acoustic and speaker variability for this kind of limited training data scenario we have investigated data augmentation methods. Firstly, we explored prosody based data augmentation and found improvement in system performance. Secondly, characteristics of Tacotron 2 was demonstrated in two ways: first training synthesis by introducing multi-speaker variability as acoustic diversity, and later by adding new utterances using language model as lexical diversity. Each method was experimented alone or through a combination which shows more performance improvement. System performance has been obtained on general and domain-specific systems through Tacotron 2 model. The best system provided a relative improvement of 50.10% with AD and 57.40% with LD method on LMS over the baseline system.

References

1. A. Sriram, H. Jun, Y. Gaur, and S. Satheesh, "Robust speech recognition using generative adversarial networks," in *Proc. ICASSP*. IEEE, 2018, pp. 5639–5643.
2. G. Evermann, H. Y. Chan, M. J. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, and P. C. Woodland, "Development of the 2003 cu-htk conversational telephone speech transcription system," in *Proc ICASSP*, vol. 1. IEEE, 2004, pp. I–249.
3. A. Ragni, K. Knill, S. P. Rath, and M. Gales, "Data augmentation for low resource languages," 2014.
4. B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," *arXiv preprint arXiv:1604.02201*, 2016.
5. J. Gu, Y. Wang, Y. Chen, K. Cho, and V. O. Li, "Meta-learning for low-resource neural machine translation," *arXiv preprint arXiv:1808.08437*, 2018.
6. S. Ren, W. Chen, S. Liu, M. Li, M. Zhou, and S. Ma, "Triangular architecture for rare language translation," *arXiv preprint arXiv:1805.04813*, 2018.
7. M. Müller, S. Stüker, and A. Waibel, "Language adaptive dnns for improved low resource speech recognition." in *Proc. INTERSPEECH*, 2016, pp. 3878–3882.
8. N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 309–314.
9. V. Kadyan, S. Shanawazuddin, and A. Singh, "Developing children's speech recognition system for low resource punjabi language," *Applied Acoustics*, vol. 178, p. 108002, 2021.
10. M. Müller and A. Waibel, "Using language adaptive deep neural networks for improved multilingual speech recognition," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, 2015.
11. A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.
12. H. Kumar Kathania, Sudarsana Reddy Kadiri, P. Alku, and M. Kurimo, "Study of formant modification for children asr," in *Proc. ICASSP*, 2020, pp. 7429–7433.
13. K. Wang, J. Zhang, S. Sun, Y. Wang, F. Xiang, and L. Xie, "Investigating generative adversarial networks based speech dereverberation for robust speech recognition," *arXiv preprint arXiv:1803.10132*, 2018.
14. T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
15. S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar, "In-domain and out-of-domain data augmentation to improve children's speaker verification system in limited data scenario," in *Proc. ICASSP*. IEEE, 2020, pp. 7554–7558.
16. H. Kathania, M. Singh, T. Grósz, and M. Kurimo, "Data augmentation using prosody and false starts to recognize non-native children's speech," in *Proc. INTERSPEECH 2020*, 2020, p. To appear.
17. J.-X. Zhang, L.-J. Liu, Y.-N. Chen, Y.-J. Hu, Y. Jiang, Z.-H. Ling, and L.-R. Dai, "Voice conversion by cascading automatic speech recognition and text-to-speech synthesis with prosody transfer," *arXiv preprint arXiv:2009.01475*, 2020.
18. S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *Proc ICASSP*. IEEE, 2018, pp. 4909–4913.
19. J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*. IEEE, 2018, pp. 4779–4783.

20. N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6706–6713.
21. Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, 2019, pp. 3171–3180.
22. W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. M. Schwartz, "Two-stage data augmentation for low-resourced speech recognition." in *Proc. INTERSPEECH*, 2016, pp. 2378–2382.
23. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.
24. C. Du and K. Yu, "Speaker augmentation for low resource speech recognition," in *Proc. ICASSP*. IEEE, 2020, pp. 7719–7723.
25. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech recognition toolkit," in *Proc. ASRU*, December 2011.
26. G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 20, no. 1, pp. 30–42, 2012.
27. D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. INTERSPEECH 2018*, B. Yegnanarayana, Ed. ISCA, 2018, pp. 3743–3747.
28. S. P. Rath, D. Povey, K. Veselý, and J. Černocký, "Improved feature processing for deep neural networks," in *Proc. INTERSPEECH*, 2013.
29. G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*. IEEE, 2013, pp. 55–59.
30. A. Stolcke, "SRILM - an extensible language modeling toolkit," in *INTERSPEECH 2002*, J. H. L. Hansen and B. L. Pellom, Eds. ISCA, 2002.
31. X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-time signal estimation from modified short-time fourier transform magnitude spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, 2007.
32. A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 996–1002.